SAMPLING FOR AREA ESTIMATION: A COMPARISON OF FULL-FRAME SAMPLING WITH THE SAMPLE SEGMENT APPROACH

MARILYN M. HIXSON, MARVIN E. BAUER
Purdue University

BARBARA J. DAVIS
Indiana Bell Telephone Company

ABSTRACT

The objective of this investigation was to evaluate the effect of sampling on the accuracy (precision and bias) of crop area estimates made from classifications of Landsat MSS data. Full-frame classifications of wheat and non-wheat for eighty counties in Kansas were repetitively sampled to simulate alternative sampling plans. Four sampling schemes involving different numbers of samples and different size sampling units were evaluated. The precision of the wheat area estimates increased as the segment size decreased and the number of segments was increased. Although the average bias associated with the various sampling schemes was not significantly different, the maximum absolute bias was directly related to sampling unit size.

I. INTRODUCTION

Accurate and timely crop production information is essential for planning the production, storage, transportation, and processing of grain crops, making marketing decisions, and determining national agricultural policies. Although most countries of the world gather crop production data, relatively few countries have reliable inventory systems. The synoptic view of the earth provided by satellite remote sensing, along with computer processing of the data, provides the opportunity to identify and estimate the area of crops.

The most comprehensive investigation of the use of Landsat MSS data for crop surveys has been the Large Area Crop Inventory Experiment (LACIE).[6] The purpose of LACIE was to assimilate current remote sensing technology into an experimental system and evaluate its potential for determining the production of wheat in various regions of the world. In LACIE, area estimates were made from classifications of Landsat MSS data. Yield was estimated for fairly broad geographic regions using statistical regression models developed from historical weather and wheat yield data.

For the area estimation phase of LACIE, samples, five by six nautical miles in size, were selected for analysis to represent about two percent of the agricultural land area. Segments were allocated to political units according to the historical area of wheat. The sample segments were used both for training the classifier and for aggregation to obtain area estimates. The LACIE method was generally successful in obtaining unbiased and precise area estimates. Six hundred segments were selected in the United States, and 1900 in the Soviet Union, to achieve a sampling error of two percent.

An alternative sampling plan for obtaining area estimates was used in another investigation at LARS.[1] A systematic sample of pixels spread throughout a Landsat full-frame was classified and used to make estimates, while training data were obtained separately. The classifications were performed on a county basis using every other line and every other column of Landsat data. Training statistics were developed using photointerpretation from aerial infrared photography taken along several flightlines dispersed throughout the state and were extended to counties lacking reference data, but known to have similar land use, crops, and soils. The pixel sampling approach was demonstrated to have the capability to produce unbiased and precise area estimates

This research was sponsored by the National Aeronautics and Space Administration, Johnson Space Center (Contract NAS9-14970).

for small (e.g., county) as well as large (e.g., state) geographic areas.

The goal of any estimation procedure is to obtain an accurate estimate. Bias and precision are both components of accuracy. Bias refers to the size of deviations from the true parameter, while precision refers to the size of deviations from the mean of all estimates of the parameter obtained through repeated applications of the sampling procedure.[2]

Numerous aspects of the crop inventory problem using remote sensing may affect the bias and precision of the estimates. Choices involving the spectral features to be measured, the sensor to be utilized, the timing of the crop observation, and the analysis methods used are all important aspects to be considered in the design of a remote sensing system. One consideration which has not been extensively researched is the choice of sampling method for area estimation.

II. OBJECTIVES

The overall objective of this investigation was to evaluate the effect of sampling on the accuracy of crop area estimates made from classifications of Landsat MSS data. The specific objectives were to assess the precision and bias associated with alternative sampling schemes involving different numbers of samples and different sampling unit sizes.

III. EXPERIMENTAL APPROACH

Ideally, a study of bias and precision of a sampling scheme would be conducted by sampling repetitively from the population of interest. In this case, however, the population of interest is the true distribution of crops in a state (or other region), and this truth is not generally known for large regions.

An alternative approach to actually conducting the experiment is to simulate its occurrence. Simulated data are used instead of truth and they are repetitively sampled to determine a variance. The estimates made are compared for bias not with truth, but with the mean of the distribution from which the data were generated.

The approach taken in this study is a combination of the two approaches described above. Full-frame classifications of Kansas into wheat and non-wheat made in another investigation[1] were used in this study as simulated ground truth. Eighty

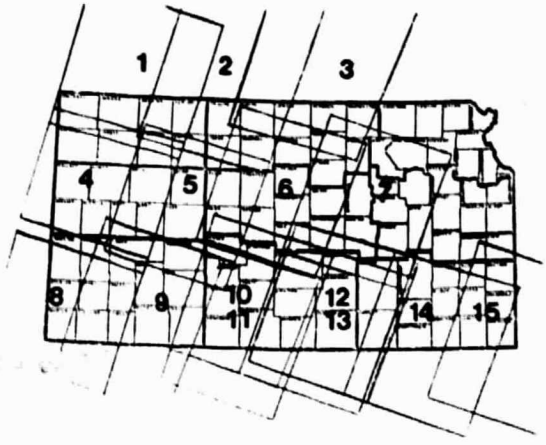

Figure 1. Landsat Full-Frame Classifications of Kansas. Alternative sampling schemes were simulated using these data.

counties comprising seven crop reporting districts were included. The Landsat frames used in these classifications are shown in Figure 1. The estimates of wheat area obtained in that study did not differ significantly from the USDA/SRS estimates at the state level. The full-frame classifications were considered to have negligible sampling error and were repetitively sampled to simulate alternative sampling plans.

Four sampling schemes were selected for testing. The total number of pixels in the sample was held constant, and the sampling unit size and number of samples were varied. Two types of samples were considered: cluster (segment) sampling and point (pixel) sampling of full-frames.

Sampling Unit Size	No. of Samples
5 x 6 nm	75
4 x 4 nm	137
2 x 2 nm	560
Pixel	427,587

Procedures similar to those followed in LACIE were used to determine the allocation (number) of samples, location (geographic placement) of segments, and the aggregated area estimate of wheat.[5,7]

A. SAMPLE SEGMENT ALLOCATION

Based on 84 sample segments which were allocated to the state of Kansas in LACIE, the number of segments per county was computed. The threshold value for each

county was computed based on the total number of acres in the county and the standard deviation of the proportion of wheat in that county. For county k,

$$t_k^* = A_k \sqrt{p_k(1-p_k)},$$

where A_k is the total land area in county k, and p_k is the historical proportion of wheat in county k. The proportional number of sample segments allotted to each county was computed by:

$$N_k = \frac{84 t_k^*}{\sum_{i=1}^{n} t_k^*}$$

where t_k^* is as defined above, and n is the number of counties in the state. The number of sample segments allotted to each crop reporting district (CRD) in the state was computed similarly.

The type of sample was then determined by the following procedure:

1. Stratified sample segment - all counties with $N_k \geq 0.5$ will have at least one sample segment; the actual number of segments is the rounded value of N_k.
2. No sample segments allotted if $N_k \leq 0.1$.
3. Probability proportional to size (PPS) sampling is done otherwise, spreading remaining segments for the CRD among the remaining counties.

Allocations strictly according to the LACIE procedure produced county allocations which did not add to the total number allocated for the crop reporting district. It was found that LACIE had also encountered this problem and had adjusted its allocations to achieve consistency. Determination of the number of segments per county followed the scheme given below for 5 x 6 nm segments because more consistent results were obtained than with the method given in the LACIE documentation:

Value of n_k	Segments Allocated
0.0 - 0.3	0
0.3 - 0.6	PPS
0.6 - 1.6	1
1.6 - 2.6	2

Two counties received two sample segments, seven counties received no sample segments, and the remainder of the counties received one segment in the 5 x 6 nm segment allocation. The criteria were generalized for other segment sizes.

B. SAMPLE SEGMENT LOCATION

The selection of sample segments was computer-implemented. This allowed a large number of segments to be chosen with little personnel time and also facilitated choice of any segment size or number of segments. The greater number of samples which could be taken through automated selection permitted statistical tests of precision. The description of the procedure which was implemented follows.

A grid, spaced six nautical miles in the east-west direction and five nautical miles in the north-south direction, was defined to cover the state of Kansas. To select a sample for a given county, the number of segments whose centers were inside the county boundaries but which did not fall entirely in the defined nonagricultural areas was determined and a sample was randomly selected from these.

The selected segment was then checked against a set of constraints. The constraints for the 5 x 6 nm segments are given here. The new segment was discarded if there was another sample segment within a 12 x 10.5 nm rectangle centered about the new segment. Then two extended rectangles were defined: one, running in the east-west direction, was 10.5 x 80 nm, and the other, running north-south, was 12 x 100 nm. Only four sample segments were permitted to fall in the east-west extended rectangle, and no more than eight sample segments were permitted to fall in the north-south extended rectangle. If the new segment caused any of these constraints to fail, it was discarded, and a new random draw was made.

Table 1. Location Constraints for the Different Segment Sizes.

Segment Size	Rectangle Considered	Segments Allowed in Extended Rectangle	
		E-W	N-S
(nm)	(nm)		
5 x 6	10.5 x 12	4	8
4 x 4	8.4 x 8	6	10
2 x 2	4.2 x 4	12	20

The location of sample segments differed in two respects from the location of the LACIE segments: first, in the definition of nonagricultural areas and second, in the number of segments permitted in a

window or extended rectangle about a given segment.

Nonagricultural areas of at least 2 x 2 miles in size were excluded from consideration as sample segments. The boundaries of urban areas, federal lands, reservoirs, etc., appearing on county maps prepared by the State Highway Commission of Kansas, Department of Planning and Development were found using a coordinate digitizer. The boundary definitions of nonagricultural areas were somewhat more crude than those defined by LACIE. The reasons for this include: (1) constraints of time (including computer time) and resources (including detailed maps) and (2) the belief that only major nonagricultural areas needed to be excluded because experience in another investigation[1] indicated that even when few nonagricultural areas are excluded, estimates of high accuracy can be obtained. The constraint that a sample segment not fall within a nonagricultural area was ignored with the pixel sampling method due to excessively high costs of computer checking for each of the nearly four million samples.

The constraints concerning the number of segments permitted in a given size rectangle centered about the sample segment and its east-west and north-south extensions to 80 nm and 100 nm, respectively, were adjusted by number and size of the rectangle to be relatively consistent with the constraints for the LACIE 5 x 6 nm segments (Table 1). This type of constraint was not feasible to use for the pixel selection procedure.

C. AREA ESTIMATION PROCEDURE

Wheat area estimates were calculated for each replication for the counties and were aggregated to obtain estimates for the crop reporting districts and state. For each crop reporting district, the area estimate was computed by

$$A_j = A_{1j} + A_{2j} + A_{3j}$$

where A_{1j} is the estimate of the area in the counties within the crop reporting district which had no segments allocated; A_{2j} is the estimate for those counties which were allocated segments with probability proportional to size; and A_{3j} is the estimate for counties allocated one or more segments.

For the m_j counties falling into class 3, A_{3j} is simply the sum of the areal proportion of wheat in each county as estimated from the sample segments multiplied by the area of the counties containing the segments:

$$A_{3j} = \sum_{k=1}^{m_j} \hat{p}_{jk} A_k$$

where $\hat{p}_{jk}$ is the wheat areal proporation in in the k^{th} county estimated from the segments and weighted according to the nonagricultural area, and A_k is the total land area in the k^{th} county.

For that set of counties in a crop reporting district to which segments were allocated with probability proportional to size, the area of wheat was estimated by:

$$A_{2j} = A_2 \frac{p_j}{m_j} \sum_{k=1}^{m_j} \frac{\hat{p}_{jk}}{p_{jk}}$$

where m_j is the number of sample segments in this set of counties; A_2 is the total land area of counties in the group; $\hat{p}_{jk}$ is the Landsat estimate of wheat proportion in the k^{th} county; p_{jk} is the agricultural census wheat proportion in the k^{th} county; and p_j is the census estimate for all counties in that group.

For the m_j counties in the j^{th} district which received no sample segments, the area estimate is:

$$A_{1j} = \frac{(A_{2j} + A_{3j})}{A_2 + A_3} x_j$$

where x_j is the agricultural census wheat area for the counties in this group, and A_i is the total land area for all counties in group i.

For each sampling plan, a standard deviation was computed for the estimate using four replications. Two sampling errors per plan and eight means per plan were available for statistical analysis. The analyses were performed using nonparametric techniques since the nonhomogeneous variances did not satisfy the requirements for classical statistical testing.

Table 2. Comparison of Bias and Precision Associated with Different Sampling Schemes.

Sampling Scheme							
Number of Samples	Sample Unit Size	Mean	Bias		Average Relative Difference	Standard Deviation	Coefficient of Variation
			Maximum	Average			
		(000 Ha)	(000 Ha)	(000 Ha)	(%)	(000 Ha)	(%)
75	5 x 6 nm	5550.9	498.2	127.5	2.4	223.7	4.0
137	4 x 4 nm	5365.0	-227.4	-58.4	1.1	86.3	1.6
560	2 x 2 nm	5409.6	80.5	-13.8	0.3	55.2	1.0
427,587	Pixel	5405.9	-39.1	-17.5	0.3	12.1	0.2

IV. RESULTS AND DISCUSSION

The effects of varying sampling unit size and the number of samples are illustrated in Figure 2 and are summarized in Table 2. Qualitative and quantitative discussions of the precision and bias of the estimates follow.

A. PRECISION

The results in Figure 2 show that the use of larger sample unit sizes results in a greater range and more variability in the estimates. The standard deviations obtained range from 11,300 hectares for pixel samples to 237,500 hectares for 5 x 6 nm segments (Table 2). Coefficients of variation range from 0.2% for pixel samples to 4.0% for 5 x 6 nm segments. The variability associated with the pixel samples is thus nearly negligible, while the 4% variability associated with one group of the 5 x 6 nm segments does not seem to be negligible.

These observations are supported by statistical results. A distribution free multiple comparison test based on the Kruskal-Wallis rank sums was performed.[4] This test was used to assess which pairs of sample unit sizes, if any, had significantly different sampling errors. At the 5% level of significance, the only pair of sampling unit sizes which had significantly different standard deviations was the 5 x 6 nm and pixel samples.

B. BIAS

The results presented in Figure 2 indicate that there may be some difference in the means of estimates made using the different sampling units. The means range from 5,365,000 hectares to 5,550,900 hectares (Table 2). Unlike the standard deviations, the means are not ranked in order according to the sample unit size.

The horizontal line in Figure 2 represents the total number of hectares of wheat in the classifications which were sampled. This number is the true population parameter which is to be estimated. A large systematic bias is not indicated since the population parameter falls in the center portion of the range of the

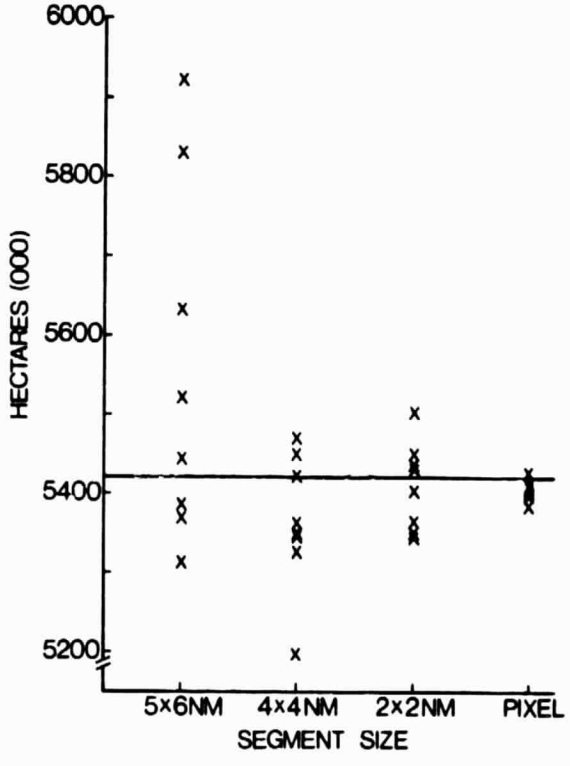

Figure 2. Comparison of Estimates Associated with Different Sampling Schemes with the Population Parameter (Horizontal Line).

estimates for all the sampling schemes, rather than most of the observations being either above or below the line. However, as indicated in Table 2, the smaller sampling units tend to yield estimates which have less bias. The average relative difference of pixel samples and 2 x 2 nm samples from the population parameter was only 0.3%, while the 5 x 6 nm segments gave estimates with an average relative difference of 2.4%.

Two types of nonparametric tests were performed to assess the bias of the several sampling methods. The Kruskal-Wallis rank sum test for one-way classifications was used to determine the effect of sampling unit size on the area estimates.[4] No significant difference in the means was found. The sign test was performed on the estimates to determine if the mean of any of the sampling schemes was significantly different from the true area of the data sampled.[3] Again, no statistically significant differences were found.

Although none of the sampling schemes appeared to have a systematic bias, it is important to examine the maximum bias which was generated by each of the sampling schemes. The maximum bias was directly related to the sampling unit size. The maximum absolute bias for pixel samples was only about 39,000 hectares, while one 5 x 6 nm sample gave an overestimate of 498,000 hectares.

In summary then, although no systematic bias is present, it is important to consider the maximum bias or range of estimates which would be obtained using a given sampling scheme in an operational setting. In practice, sampling would be conducted only once; thus, a one in eight chance of obtaining a bias of 500,000 hectares may be a significant consideration.

V. SUMMARY AND CONCLUSIONS

The results of this investigation are well illustrated in Figure 2. The area estimates found by the use of 5 x 6 nm segments cover a much larger range of values and thus have a larger variability than any of the other segment sizes. The estimates become more and more precise as the segment size decreases and more segments are taken. The estimates achieved using the 5 x 6 nm segments have the least precision of any sampling scheme tested. The precision of the 5 x 6 nm segments was significantly less than that of the pixel samples.

None of the sampling schemes was significantly biased on the average, and none of the average estimates differed significantly from the population parameter. The maximum absolute bias, however, was directly related to sampling unit size and should be considered in selection of a sampling unit.

To assess the implications of the result of this study for operational use, other factors must be considered. In order to fully evaluate the scheme, the method of training and classification which would be used in conjunction with a sampling plan must also be considered. And, although the precision of estimates from choosing more but smaller segments may be higher, this gain in precision must be weighed against the costs of sample selection and classification.

A somewhat similar study was recently conducted by Perry.[8] The objective of that study was to ascertain the effect of a change in the sampling unit size on the total number of sampling units necessary to support a wheat production estimate with a specified coefficient of variation. The results obtained by Perry are supportive of the conclusions of this investigation, but it was concluded that no recommendation for the optimal sampling unit size can be made until a model for the cost as a function of the sampling unit size is developed.

VI. REFERENCES

1. Bauer, Marvin E., Marilyn M. Hixson, Barbara J. Davis, and Jeanne B. Etheridge. 1978. Area Estimation of Crops by Digital Analysis of Landsat Data. Photogrammetric Engineering and Remote Sensing, 44:1033-1043.

2. Cochran, W.G. 1963. Sampling Techniques. Second Edition. John Wiley and Sons, Inc., New York.

3. Freund, John E. 1962. Mathematical Statistics. Prentice-Hall, Inc., Englewood Cliffs, N.J.

4. Hollander, Myles and Douglas A. Wolfe. 1973. Nonparametric Statistical Methods. John Wiley and Sons, Inc., New York.

5. LACIE Staff. 1974. Crop Assessment Subsystem Requirements, Level III Baseline. LACIE-00200, Volume IV, pp. B1-B6. NASA, Johnson Space Center, Houston Texas.

6. MacDonald, R.B. and F.G. Hall. 1978. "LACIE: An Experiment in Global Crop Forecasting". Proceedings of Plenary Session, LACIE Symposium. October 23-26, 1978. NASA, Johnson Space Center, Houston, Texas, pp. 17-48.

7. MacDonald, R.B., F.G. Hall, and R.B. Erb. 1975. "The Use of Landsat Data in Large Area Crop Inventory Experiment (LACIE)". Proceedings, Symposium on Machine Processing of Remotely Sensed Data. June 3-5, 1975. Purdue University, West Lafayette, Indiana, pp. 1B-1 to 1B-23.

8. Perry, C.R. "Sampling Unit Size vs. Variance". Presentation at SR&T Quarterly Review. March 6-9, 1979. NASA, John Space Center, Houston, Texas.

Marilyn M. Hixson, research statistician in LARS' Crop Inventory Systems Research; B.S. in mathematics from Miami University; M.S. in mathematical statistics from Purdue University. Ms. Hixson's work at LARS has involved experiment design, data analysis, stratification, and sampling methodology. She has had a major role in the design, Landsat data classifications, and statistical analysis of results in several Landsat investigations concerning training, classification, and area estimation procedures for crop inventory.

Marvin E. Bauer, research agronomist and program leader of LARS' Crop Inventory Systems Research; B.S.A. and M.S., Purdue University; and Ph.D., University of Illinois. Dr. Bauer has had key roles in the design, implementation, and analysis phases of several major remote sensing projects including the 1971 Corn Blight Watch Experiment and the Crop Identification Technology Assessment for Remote Sensing Project. He has been the principal investigator of a Landsat investigation for crop area estimation survey. Currently, he is the technical leader of the agricultural field research program at LARS.

Barbara J. Davis, B.S., mathematics, Michigan State University; M.S., applied statistics, Purdue University. She was a Statistician/Analyst at LARS from 1973 to 1978, including the period in which this work was done. Her work at LARS included algorithm development, crop inventory surveys, and the application of statistical methods to remote sensing problems.

Mrs. Davis is currently a Staff Associate in Business Research at Indiana Bell Telephone Company in Indianapolis, where she designs and conducts surveys and provides statistical consultation for all departments of the company. She is a member of the Central Indiana chapter of the American Statistical Association.